Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт космических и информационных технологий
институт
Вычислительная техника
кафедра

УТВЕРЖДАЮ
Заведующий кафедрой
_____ О.В.Непомнящий
подпись инициалы, фамилия
« ___» _____ 2018 г.

БАКАЛАВРСКАЯ РАБОТА

09.03.01 Информатика и вычислительная техника
код и наименование направления

Разработка программного обеспечения для анализа многомерных данных ядерным методом главных компонент
тема

Руководитель	_____	доцент, канд.техн.наук	Л. И. Покидышева
	подпись, дата	должность, ученая степень	инициалы, фамилия
Выпускник	_____		П. А. Толкачёв
	подпись, дата		инициалы, фамилия
Нормоконтролер	_____	доцент, канд.техн.наук	В. И. Иванов
	подпись, дата	должность, ученая степень	инициалы, фамилия

Красноярск 2018

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В любое из времен человек совершенствуется и улучшает технологии, которыми пользуется. Многие исследователи, занимающиеся анализом многомерных данных, сталкиваются с проблемами их интерпретации и классификации, и извлечения другой полезной информации. Для решения таких задач существуют специальные методы многомерного анализа. Важным критерием выбора метода является потеря информации при уменьшении размерности, а также на выбор влияет тип и размерность данных.

Методы многомерного анализа находят широкое применение в медицине, статистике, психологии, экономике и других науках. Данные методы используются для визуализации данных, подавления шума на изображениях, психодиагностики, распознавания образов и для решения множества других задач. Многомерные методы уменьшения размерности позволяют визуализировать данные. В этом случае, такие методы используются для отображения на двумерное или трехмерное пространство с последующим выделением кластеров, наглядным анализом взаимного расположения объектов, выделением их общих характеристик.

Проблемой уменьшения размерности многомерных данных занимаются многие ученые. На протяжении последних лет создано множество алгоритмов, методов и подходов, решающих описанные выше задачи. Каждый из методов имеет свои преимущества и недостатки, они опираются на различные математические принципы. Многие исследователи изучают возможности различных методов в прикладных сферах для решения конкретных задач, описывают их преимущества и недостатки. Во многих задачах обработки многомерных наблюдений и, в частности, в задачах классификации исследователя интересуют в первую очередь лишь те признаки, которые обнаруживают наибольшую изменчивость (наибольший разброс) при переходе от одного объекта к другому.

Особое внимание проблемам многомерного анализа данных уделяют разработчики математических библиотек. Зачастую целые институты и другие сообщества поддерживают библиотеки в актуальном состоянии, дополняют их новыми алгоритмами, исправляют ошибки, снабжают примерами и т.д. Платные математические пакеты предоставляют только базовые методы анализа и, в большинстве случаев, исследователю необходимо реализовывать алгоритмы самостоятельно на каком-либо языке программирования, что не всегда удобно.

1 Анализ задания на выпускную квалификационную работу

Целью работы является разработка программного обеспечения для анализа многомерных данных ядерным методом главных компонент.

Реализация поставленной цели предполагает необходимость решения следующих задач:

– изучение классического и ядерного методов главных компонент;

– изучение существующих аналогичных программных средств обработки данных;

– выбор инструментария и способов решения;

– разработка программного обеспечения, реализующего ядерный метод главных компонент.

1.1 Требования, предъявляемые к системе

Программное обеспечение для анализа многомерных данных ядерным методом главных компонент должен:

– иметь удобный, интуитивно понятный интерфейс;

– предоставлять пользователю помощь в работе с системой;

– выполнять не только анализ многомерных данных, но и предварительную обработку;

– работать с данными, представленными в формате CSV;

– сохранять данные в формате CSV;

– работать с большими массивами данных;

– сохранять результаты анализа в виде графика;

– иметь возможность расширения с целью добавления новых методов.

1.2 Обзор существующих аналогичных систем

1.2.1 ViDaExpert

Разработка программного обеспечения ViDaExpert началась в 2000 году Андреем Зиновьевым, выпускником СФУ (Россия). На данный момент программа активно используется исследователями Института Кюри (Франция), одного из лидирующих научных институтов в области биофизики, молекулярной биологии и онкологии [1].

ViDaExpert реализует ряд методов анализа исследовательских данных:

– анализ основных компонентов;

– анализ взвешенных компонентов;

– иерархическая кластеризация;

– линейный регрессионный анализ;

– линейный дискриминантный анализ и др.

На рисунке 1 приведен пример интерфейса ViDaExpert в процессе работы с данными.

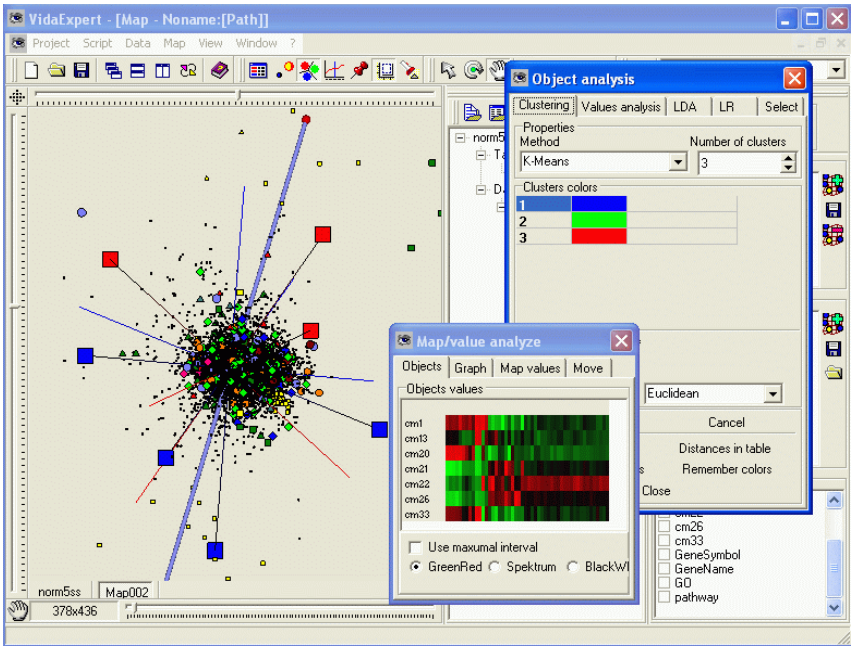

Рисунок 1 – Интерфейс ViDaExpert в процессе работы с данными

В ViDaExpert имеется хорошо разработанный набор инструментов для просмотра, аннотации и маркировки данных с цветами, формами и размерами.

Но для простого пользователя программа слишком сложна и многофункциональна.

1.2.2 MATLAB

MATLAB предоставляет инструменты для получения, анализа и визуализации данных, позволяющие исследовать проблему быстрее, чем это возможно с помощью электронных таблиц или программирования.

MATLAB позволяет получать доступ к данным из файлов, других приложений, баз данных, внешних устройств. Имеет возможность читать данные из файлов таких форматов как Microsoft Excel, текстовых или двоичных файлов, изображений, аудио и видео файлов, научных форматов (netCDF и HDF). Функции ввода-вывода позволяют работать с файлами данных любых форматов.

MATLAB позволяет управлять, фильтровать и осуществлять предварительную обработку данных. Позволяет исследовать данные для нахождения трендов, проверки гипотез, построения описательных моделей. В MATLAB включены функции для фильтрации, сглаживания, свёртки и быстрого преобразования Фурье (FFT). Продукты-расширения включают возможности подбора кривых и поверхностей, многомерной статистики, спектрального анализа, анализа изображений, идентификации систем и другие инструменты анализа.

MATLAB предоставляет набор встроенных функций построения 2D и 3D графиков, а также функции объёмной визуализации [2].

На рисунке 2 представлена реализация метода главных компонент в программе MATLAB.

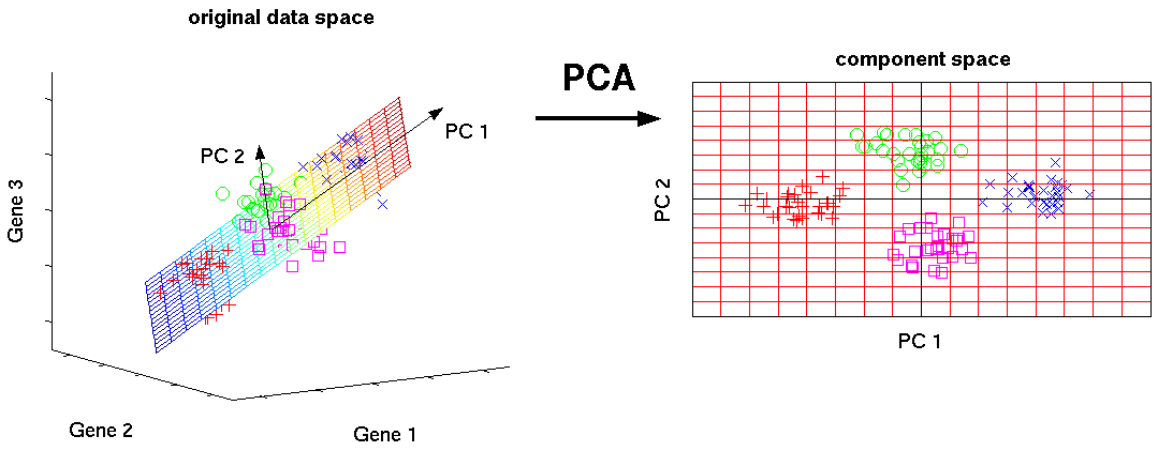

Рисунок 2 – Реализация метода главных компонент в программе MATLAB

1.2.3 Программное обеспечение по статистическому анализу данных: методология сравнительного анализа и выборочный обзор рынка ПО

Класс инструментальных средств поддержки процессов анализа данных – объединеняет одним свойством входящие в него средства: все они направлены на преодоление проблемы большой размерности. В результате появляется необходимость снижения размерности, выделения именно тех фрагментов данных, которые представляют интерес для решения конкретной проблемы. Для пользователей, имеющих дело со сверхбольшими объемами данных, характеризующихся высоким уровнем формализации представления, серьезной альтернативы использованию этого класса программного обеспечения пока нет.

На сегодняшний день лидирующие позиции занимают западные пакеты статистической обработки и среды математического моделирования. В большинстве из них реализованы специальные высокоуровневые языки программирования для реализации собственных алгоритмов обработки данных. Их разработка осуществляется путем комбинирования готовых подпрограмм, поставляемых с данным программным продуктом в специализированных библиотеках. Пользователям предоставляется возможность разработки собственных процедур с применением встроенных средств разработки или внешних сред программирования. Универсальные пакеты обладают несколько меньшими возможностями, но их стоимость значительно ниже, чем стоимость

профессиональных. Универсальные пакеты схожи по составу методов обработки, реализованы по модульному принципу и за счет обращения к процедурам и функциям операционной системы упрощают работу с графикой и интерфейсными элементами. Однако, при приобретении таких систем следует убедиться, что они действительно реализуют требуемые методы и алгоритмы обработки данных.

Существует около тысячи распространяемых на мировом рынке пакетов, решающих в том или ином виде задачи статистического анализа данных, и реализованные для различных операционных систем.

В СНГ также интенсивно развивается направление, связанное с разработкой программного обеспечения для статистической обработки данных. К этому направлению могут быть отнесены российские пакеты STADIA103 (НПО «Информатика и компьютеры»), ОЛИМП (ЗАО «CPS») и белорусский пакет РОСТАН (Белорусский Государственный Университет). Имеются примеры создания специализированных систем для решения задач классификации и снижения размерности, например: КЛАСС МАСТЕР (Научное изд-во «ТВП»), КВАЗАР (ИММ УрО РАН), PALMODA (ВЦ РАН), Stat-Media (ЗАО «Полихимэкс») и иные. Кроме того, на рынке представлены и статистические экспертные системы, например, СТАТЭКС (РМ и ПК, Казахстан). Довольно интересный класс программного обеспечения представляют собой системы, ориентированные на решение задач снижения размерности, классификации и анализа данных. Эти системы используют комбинацию методов статистической и нейросетевой104 обработки данных. В этой области столь эффективно работают такие гиганты, как SAS (серия продуктов SAS Data and Text Mining) и SPSS, создающие программные продукты, сочетающие мощь статистических методов обработки с методами нейрокомпьютинга. Среди наших разработчиков следует отметить ВЦ РАН (ЛОРЕГ), ЗАО «Megaputer» (система PolyAnalyst), НПИЦ «Микросистемы» (система TextAnalyst), фирму «Контекст» (пакет «ДА-система») и «MediaLingua» (система Классификатор) [3].

В медицине это может быть диагностика состояния пациента по комплексу наблюдаемых признаков (результаты клинического осмотра, лабораторных исследований, оцифровки и кодирования рентгенограммы и/или сонограммы). В геофизике – прогноз степени перспективности месторождения нефти или газа, в области финансов – оценка уровня кредитоспособности клиента или прогноз тенденции поведения рынка ценных бумаг, в экономике – разнообразные задачи типологизации объектов (семей, предприятий, городов, стран и т.п.) и прогноза социально-экономического поведения «хозяйствующего субъекта», в маркетинге – позиционирование нового товара среди существующих, в технике – диагностика состояния турбины или двигателя, контроль уровня качества продукции и др.

В таблице 1 представлены общие сведения об универсальных пакетах и сведения о минимальных аппаратных требованиях к ним для операционных систем Windows и MS-DOS.

Таблица 1 – Общие сведения об универсальных пакетах и сведения о минимальных аппаратных требованиях к ним

Стат.Система	Версия	ОС	МП/Част.	VHD	RAM	Фирма-продавец	User	Цена
SAS	6.11 6.07	W D	386/33	65* 44	8*** 4	SAS Institute, Inc.	H	850
Statgraphics+ Statgraphics+ Statgraphics	1.0 7.0	W D D	386/33 286/12	14.5 8.5 6.1	4 4 1	Manugistics, Inc.	M-L M M	1048 995 995
MINITAB	10.0 7.0	W D	386/16 286/12	12 4	4 1	MINITAB Inc.	M-L	895
SYSTAT	6.0 6.0	W D	386/33	8	4	SPSS, Inc.	H	995 995
SPSS/PC	7.0	W	486/50	65**	8	SPSS, Inc.	H	980
BMDP Dynamic		D				SPSS, Inc.	H	695
STATISTICA	5.1	W D	386/33	13	4	StatSoft, Inc.	H-M	995 795
IMSL-C (Num) Object Suite	2.0 1.0	W W				Visual Numerics StatSci	H H	700 700
S-Plus		W D					H H	1450 1195

Пояснения к таблице 1:

1. ОС – сокращение от «Операционная система»: W – Windows, D – MS-DOS;

2. Размеры в Мб: VHD – место, занимаемое на винчестере; RAM – операт. память;

3. МП – основной микропроцессор; Част. – его тактовая частота в [МГц];

4. User – квалификация типичного пользователя: H (high) – статистик профессионал M (middle) – «есть базовые статистические знания»; L (low) - «отсут. базового уровня»; H-M – промежуточный;

5. Цены указаны в [$]: цена лицензионной копии СПП взята из каталогов [5, 6, 7]};

6. Цены для SAS и SPSS указаны для базовых модулей на рынке в РФ; кроме того, SAS требует ежегодную оплату лицензии. С другой стороны, достаточно полная конфигурация SPSS (модуль Base + комплект из семи модулей) стоит $4290; Каждый из дополнительных модулей SAS или SPSS стоит, как правило, от $350 до $750. Цена на STSC+/W указана на комплект: «Базовый модуль» плюс «Модуль многомерного анализа».

7. * для модулей BASE, STAT, GRAPH;

** включая файл «подкачки» на диске;

*** дополнительно рекомендуется файл «подкачки» на диске размером 15 Мб.

1.3 Обзор классического и ядерного метода главных компонент

1.3.1 Классический метод главных компонент

Метод главных компонент (МГК) является основным и самым используемым методом уменьшения размерности данных.

Суть метода состоит в преобразовании исходного набора данных на пространство меньшей размерности таким образом, чтобы выполнялись следующие условия:

– минимальна сумма квадратов расстояний от точек, данных до их проекций на плоскость первых главных компонент, то есть экран расположен максимально близко по отношению к облаку точек;

– минимальна сумма искажений квадратов расстояний между всеми парами точек из облака данных после проецирования точек на плоскость;

– минимальна сумма искажений квадратов расстояний между всеми точками данных и их «центром тяжести».

На рисунке 3 вектора v1 и v2 являются главными компонентами (ортогональные проекциями), на которые отображается входной набор данных.

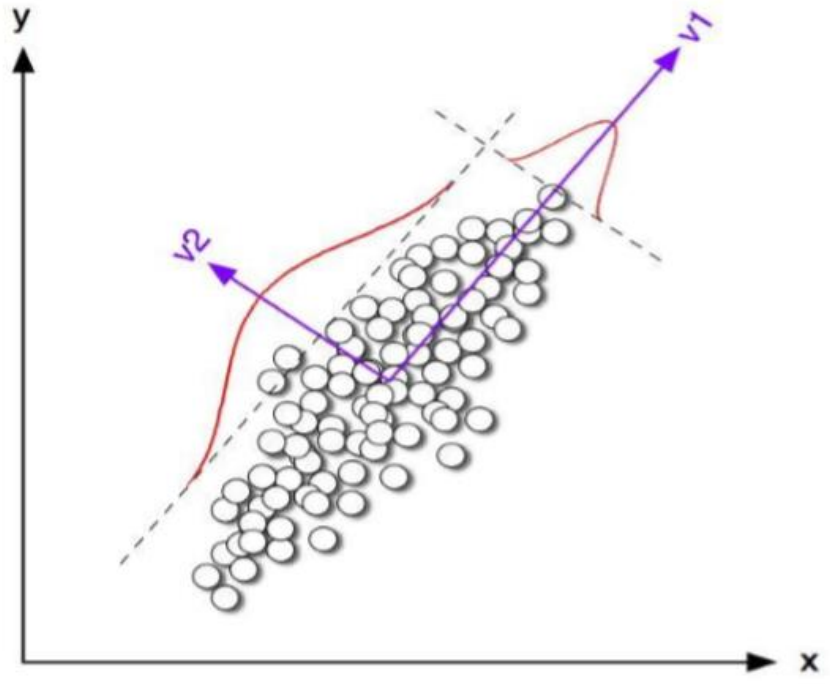

Рисунок 3 – Главные компоненты набора объектов

Главные компоненты находятся за счет вычисления собственных векторов и собственных значений ковариационной матрицы, составленной на основании входной матрицы данных. Те собственные вектора, у которых собственные значения наибольшие, описывают наибольшую дисперсию, собственный вектор с самым меньшим собственным значением – наименьшую дисперсию [1].

Метод главных компонент стремится выделить оси, вдоль которых количество информации максимально, и перейти к ним от исходной системы координат. При этом некоторое количество информации может теряться, но зато сокращается размерность. Этот метод проходит практически через весь факторный анализ, и может меняться путем подачи на вход разных матриц, но суть его остается неизменной. Основной математический метод получения главных осей – нахождение собственных чисел и собственных векторов ковариационной матрицы. Сумма собственных чисел равна числу переменных, произведение - детерминанту корреляционной матрицы. Собственное число представляет собой дисперсию оси, наибольшее - первой и далее по убыванию до наименьшего - количество информации вдоль последней оси. Доля дисперсии, приходящаяся на данную компоненту, считается отсюда легко: надо разделить собственное число на число переменных m. Коэффициенты нагрузок для главных компонент получаются делением коэффициентов собственных векторов на квадратный корень соответствующих собственных чисел.

Главным недостатком данного метода является то, что линейные преобразования данных не учитывают взаимное расположение точек в пространстве. В случае, если данные имеют определенную структуру, которую возможно использовать для наиболее точного уменьшения размерности, метод главных компонент не является эффективным.

1.3.2 Ядерный метод главных компонент

Ядерный (Kernel) МГК является одним из нелинейных методов уменьшения размерности данных. Он соединяет в себе линейный МГК и специальный набор преобразований, называемый kerneltrick.

Kerneltrick позволяет преобразовать любой алгоритм, который использует скалярное произведение между двух векторов. Для применения такого преобразования используется замена всех скалярных произведений ядерной функцией. Таким образом, можно сказать, что линейный алгоритм преобразуется в нелинейный. Этот нелинейный алгоритм эквивалентен линейному, только в пространстве большой размерности ϕ. Но суть заключается в том, что функция ϕ никогда не вычисляется напрямую. А значит пространство ϕ может быть сколь угодно велико.

Суть работы ЯМГК представлена на рисунке 4.

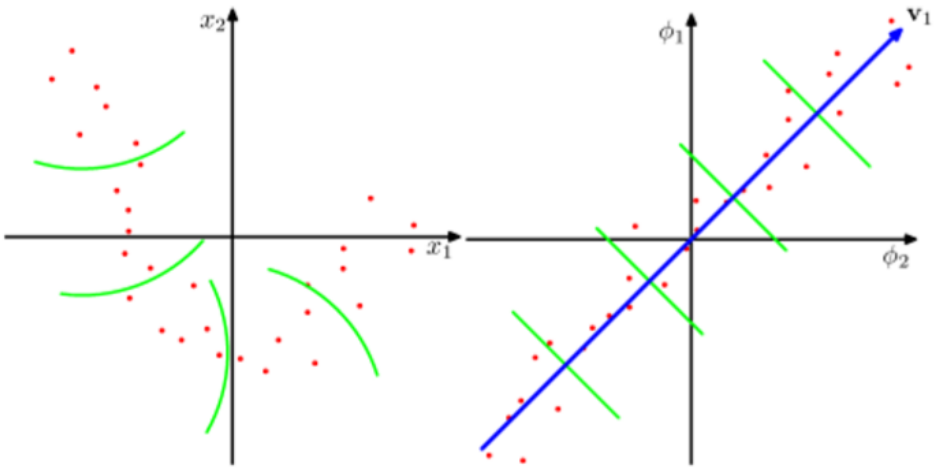

Рисунок 4 – Пример работы ядерного метода главных компонент

Как и в обычном МГК, целью преобразования является максимизация дисперсии одних признаков и одновременная минимизация ковариации среди других. В ЯМГК ковариационная матрица заменяется ядерной матрицей, а дальнейшие преобразования аналогичны линейному МГК [4].

Рассмотрим алгоритм работы более подробно. Вход алгоритма состоит из матрицы D, а также дополнительных параметров, зависящих от функции ядра.

Шаг 1. Вычисляем ядерную матрицу по формуле 1:

$$K_{ij} = l(D_i, D_j), \quad (1)$$

где D – матрица входных данных;

l – номер компоненты;

K – ядерная матрица.

Шаг 2. Центрируем и диагонализируем матрицу K по формуле 2:

$$K_C = K - 1_N K - K 1_N + 1_N K 1_N, \quad (2)$$

где 1_N – квадратная матрица размера N;

K – то же, что и в формуле (1).

Квадратная матрица размера N определяется по формуле 3:

$$(1_N)_{ij} = \frac{1}{N}, \quad (3)$$

где 1_N- то же, что и в формуле (2);

N – размер матрицы K.

Шаг 3. Нормализуем полученную матрицу K_C и находим её собственные значения и собственные вектора.

Шаг 4. Вычисляем отображение на первые k главных компонент по формуле 4:

$$F_k = \sum_{i=1}^{N} v_i^k \, l(D_i, D_j) \ , \qquad (4)$$

где l – номер компоненты (не экспонента);

F – матрица результата;

v – собственные значения;

D – то же, что и в формуле (1).

В алгоритме выше L(x,y) является одной из ядерных функций. Наиболее часто используются следующие ядерные функции: полиномиальная и Гауссова

Полиномиальная ядерная функция представлена в формуле 5:

$$l(x, y) = (ax^T y + c)^d \ , \qquad (5)$$

где a – множитель;

c – слагаемое;

d – степень;

x – вектор;

y – вектор.

Гауссова ядерная функция представлена в формуле 6:

$$l(x, y) = e^{\left(-\frac{|x-y|^2}{2\sigma^2}\right)}, \qquad (6)$$

где σ – параметр;

e – экспонента;

x – вектор;

y – вектор.

Параметры ядерных функций (слагаемые, степени и множители) задаются перед началом работы алгоритма и не изменяются во время работы.

1.3.3 Задача выбора количества главных компонент

При использовании ядерного метода главных компонент важно правильно определить оптимальное количество главных компонент. Если их число слишком мало, то описание данных будет не полным. Избыточное число главных компонент приведёт к переоценке, то есть моделируется не содержательная информация.

Для выбора количества главных компонент используется график, на котором объясненная дисперсия (ERVC) изображается в зависимости от числа главных компонент (PC), как на рисунке 5.

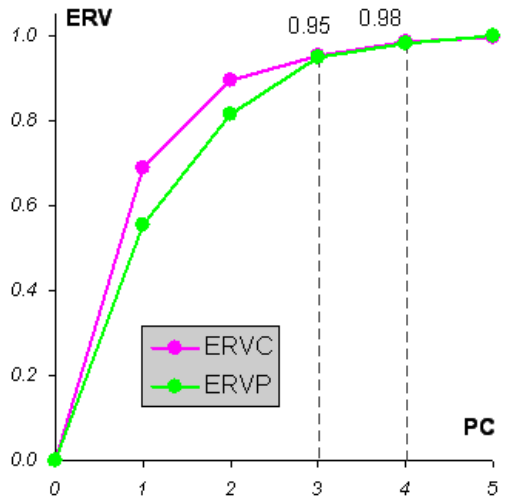

Рисунок 5 – Зависимость EVR от PC

На графике видно, что верными решением будет выбрать число главных компонент равным трём или четырём. Три компоненты отображают 95% исходной информации, а четыре – 98%.

Для реализации программной системы для анализа многомерных данных ядерным методом главных компонент данный способ будет нужным решением задачи о выборе количества главных компонент.

2 Описание программного комплекса

2.1 Структура системы

Структурная схема – это совокупность элементарных звеньев объекта и связей между ними, один из видов графической модели. Под элементарным звеном подразумевается часть объекта, системы управления и т. д., которая реализует элементарную функцию.

Для более полного понимания того, как работает система на рисунке 6 представлена структурная схема системы.

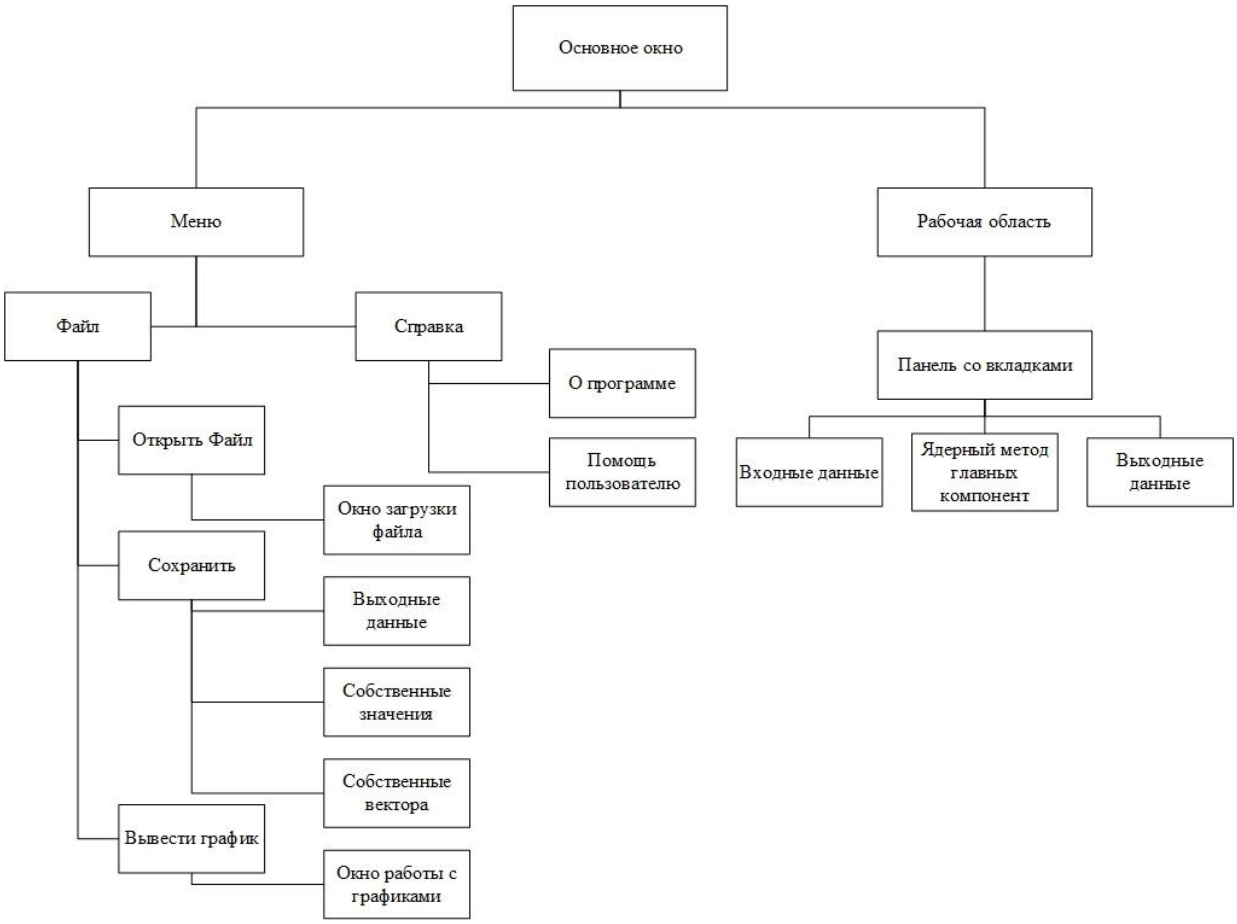

Рисунок 6 – Структурная схема системы

Основное окно программы предоставляет доступ ко всем функциям системы при помощи меню и рабочей области.

Меню представляем из себя элемент под названием «MenuStrip» из стандартного набора элементов Windows Form, содержит два пункта «Файл» и «Справка». Пункт под названием «Файл» включает в себя такие элементы как:

– «Открыть файл» – предоставляет доступ к окну загрузки файлов с данными;

– «Сохранить» – предоставляет возможность сохранить выходные данные, собственные вектора и значения в файл;

– «Вывести график» – предоставляет доступ к окну для работы с графиками.

Пункт «Справка» предоставляет доступ к информации о программе и разработчике и помощи пользователю, где содержится инструкция по пользованию системой.

Рабочая область представляет собой элемент «TabControl» который в свою очередь предоставляет доступ к трём вкладкам:

– «Входные данные» – содержит таблицу с входными данными, заполняется после загрузки входного файла;

– «Ядерный метод главных компонент» – осуществляет все расчёты для реализации метода;

– «Выходные данные» – содержит таблицу с выходными данными, заполняется после работы алгоритма реализации метода.

2.2 Разработка алгоритма системы

Алгоритм – набор инструкций, описывающих порядок действий исполнителя для достижения некоторого результата. В старой трактовке вместо слова «порядок» использовалось слово «последовательность», но по мере развития параллельности в работе компьютеров слово «последовательность» стали заменять более общим словом «порядок».

Рассмотрим общий алгоритм работы программы на рисунке 7.

Рисунок 7 – Общая схема алгоритма программы

После начала работы программы необходимо открыть файл с входными данными, после открытия файла данные, при необходимости, редактируются. Далее происходит выбор ядерной функции и количества главных компонент и

алгоритм переходит к реализации метода. На следующей стадии результаты работы метода сохраняются в файлы или в виде графиков. Если выбран выход – алгоритм заканчивает работу, иначе переходит к выбору ядерной функции и количества компонент.

Рассмотри подробнее процесс реализации ядерного метода главных компонент на рисунке 8.

Рисунок 8 – Алгоритм ядерного метода главных компонент

2.3 Описание разработанных классов

Для разработки программной системы для анализа многомерных данных ядерным методом главных компонент были разработаны классы: Method, WorkForm, Grafik. Далее данные классы будут рассмотрены подробно, включая структуры данных входящие в классы и методы.

2.3.1 Класс WorkForm

Данный класс включает в себя структуру Data, в которую будут записаны данные при работе с программой. В данной структуре будут храниться: входные и выходные данные, количество нужных главных компонент и выбранная ядерная функция.

В классе WorkForm реализованы методы для работы с основной частью рабочего интерфейса, рассмотрим наиболее важные из них:

– void vis(double[,] data) – осуществляет приём входных данных из файла и заносить в таблицу для отображения пользователю;

– private void открытьФайлToolStripMenuItem_Click(object sender, EventArgs e) – реализует открытие формы загрузки файла, при нажатии соответствующей вкладки меню;

– private void Analys_Click(object sender, EventArgs e) – вызывается при нажатии кнопки «Начать анализ», создаёт переменную класса method с помощью которого выполняется обработка данных ядерным методом главных компонент;

– void Out_data() – выводит на вкладку «выходные данные» в таблицу данные полученные в результате работы разработанного алгоритма;

– private void GrafikToolStripMenuItem_Click(object sender,EventArgs e) – открывает форму для вывода графиков визуализирующих работу ядерного метода главных компонент.

2.3.2 Класс Method

Класс Method отвечает за выполнение всех расчётов, связанных с реализацией ядерного метода главных компонент и включает в себя следующие методы:

– public double funckernel (ref double[] x, ref double[] y, ref IKernel func) – на основе входных векторов и выбранного пользователем типа ядерной функции рассчитывает каждый элемент ядерной матрицы, по формуле 4;

– double[,] mult_matrix (ref double[,] X, ref double[,] Y) – выполняет умножение матрицы X на матрицу Y по правилам перемножения матриц;

– double[,] morp_matrix(ref double[,] X, ref double[,] Y, bool what) – выполняет сложение или вычитание матриц X и Y, в зависимости от входного параметра what, если what=true, то выполняется сложение, иначе – вычитание;

– public double[,] Kc (ref double[,] K) – центрирует и диоганализирует матрицу K по формуле 5, а также вычисляет квадратную матрицу размера N по формуле 6;

– public double[,] GetV(ref double[][] eigenvect, ref double[][] eigenval, ref double[,] Kc) – вычисляет собственные значения и собственные вектора матрицы Kc, сортирует собственные вектора по собственным значениям, записывает в новую матрицу V и вычисляет отображение на определенное количество главных компонент, возвращает в рабочую форму готовые выходные данные.

2.3.3 Класс Grafik

Данный класс отвечает за вывод данных на график, для наглядного изучения. Для вывода доступны входные данные, выходные данные и собственные значения. Позволяет сделать выбор какие компоненты вывести на график и на какую ось.

Private void DrawGra_Click(object sender, EventArgs e) – метод отвечает за заполнение графика данными и за отчистку графика.

2.4 Описание графического интерфейса

Графический пользовательский интерфейс в первую очередь должен быть интуитивно понятным для любого пользователя, поэтому при разработке программной системы особое внимание уделялось реализации как можно более понятного, но в то же время функционального интерфейса.

Основное окно программы состоит из меню, 3 вкладок и одной кнопки.

Для начала работы алгоритма нужно открыть файл с данными для анализа, на рисунке 9 изображено основное окно программы и меню, где можно загрузить файл.

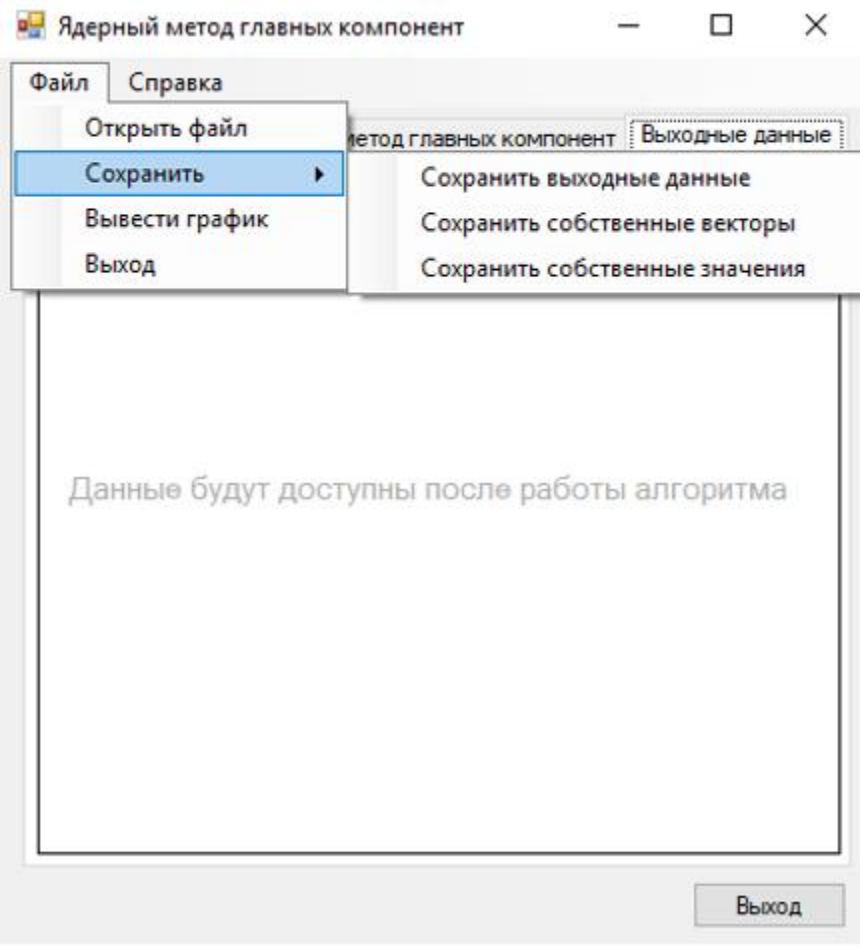

Рисунок 9 – Основное окно и меню

После выбора пункта «Открыть файл» открывается новое окно, как на рисунке 11, для работы с файлом, где можно указать путь к файлу с данными и предварительно посмотреть содержимое файла.

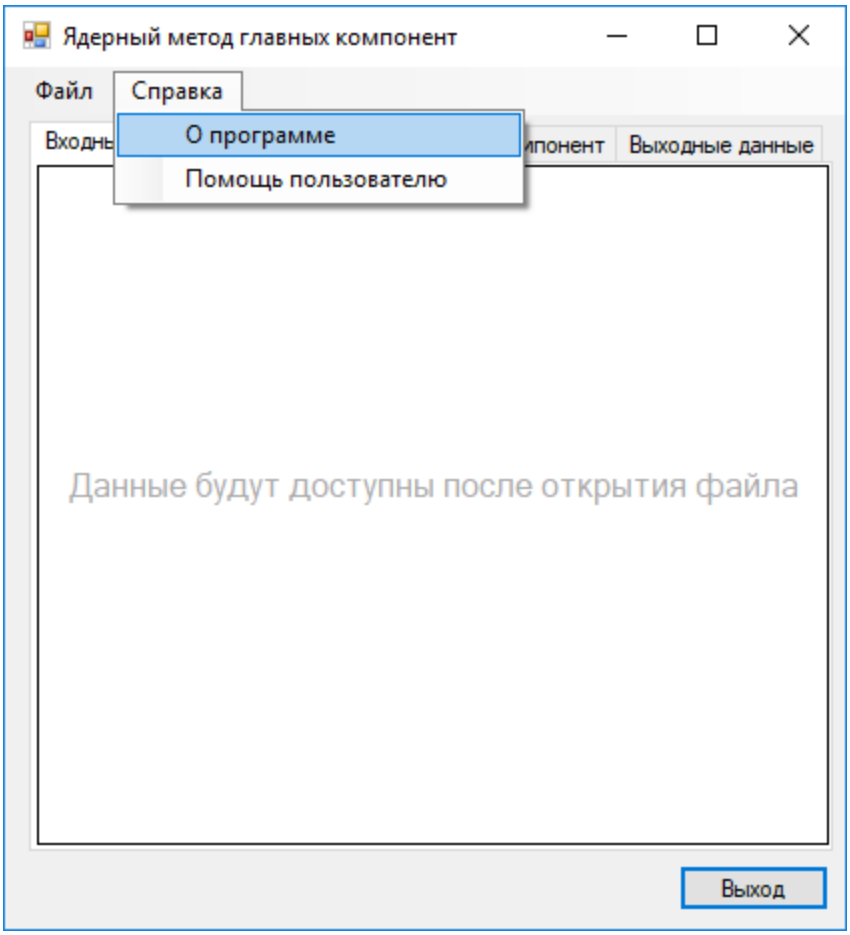

Рисунок 10 – Элемент меню «Справка»

На рисунке 10 показан пункт меню «Справка», где можно узнать всю информацию о программе, разработчике и получить инструкцию по пользованию программой.

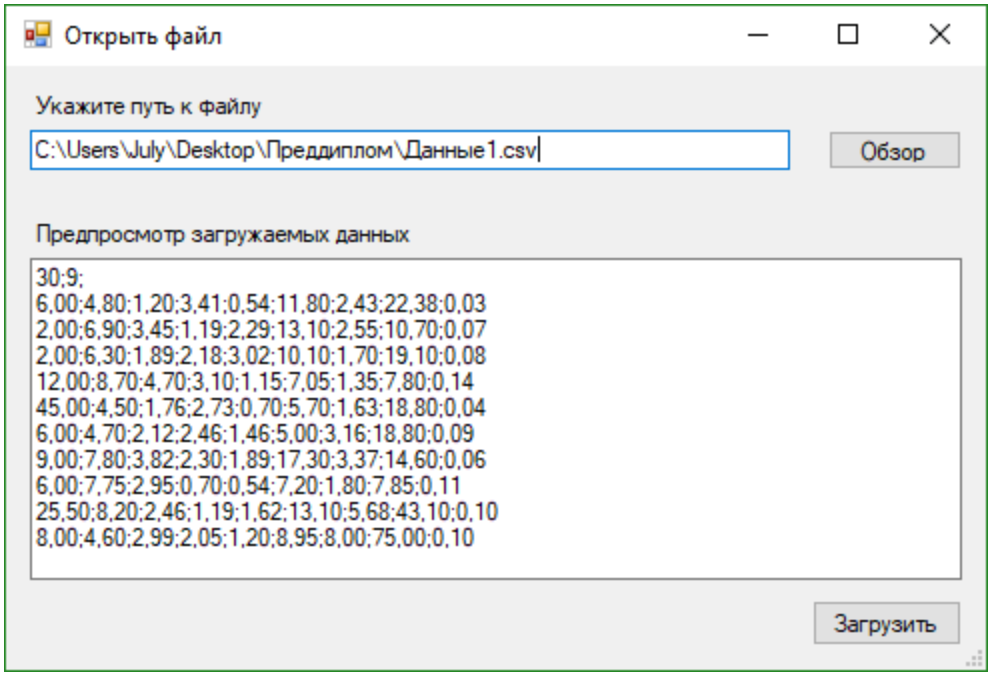

Рисунок 11 – Окно открытия файла

С помощью предпросмотра данных, пользователь может определить нужный ли файл он открывает. Числа отделяются друг о друга точкой с запятой, входной файл не должен иметь пропусков в значениях, первые два числа обозначают размерность данных.

После выбора файла и нажатия кнопки «Загрузить» открывается основное окно с таблицей входных данных, как на рисунке 12.

Теперь пользователь может редактировать данные, но не добавлять или удалять.

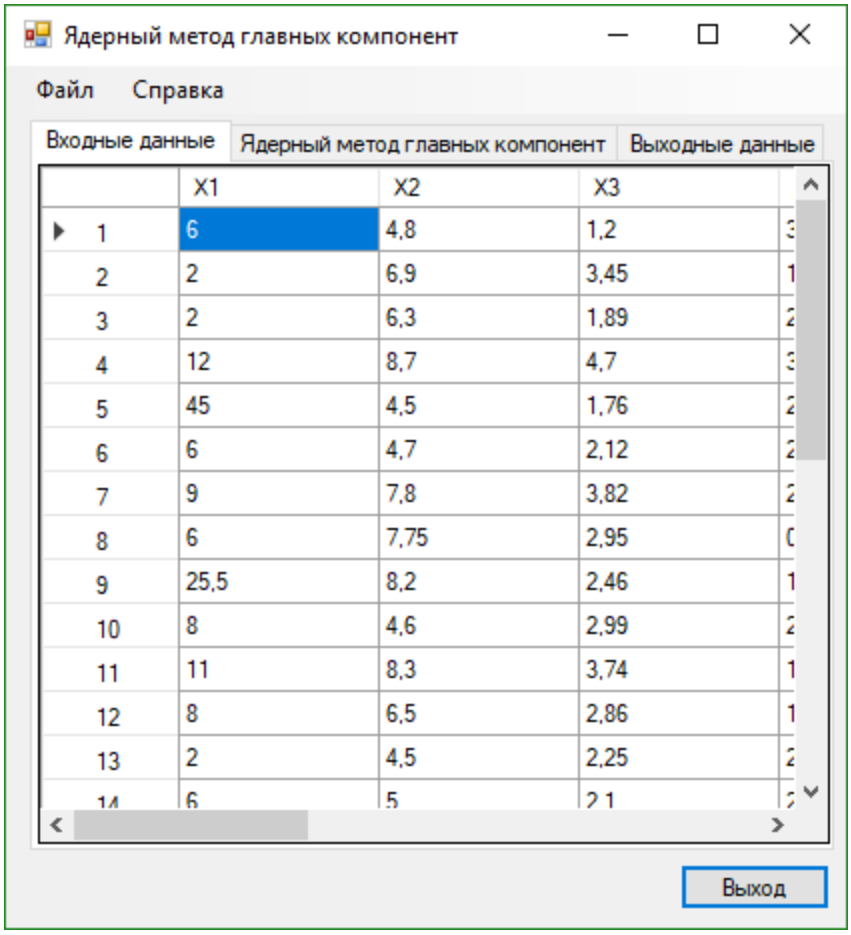

Рисунок 12 – Основное окно после загрузки файлов с таблицей входных данных

После загрузки файла с данными, при необходимости отредактированных, Пользователь может перейти во вкладку «Ядерный метод главных компонент» изображенную на рисунке 13. Здесь пользователю нужно выбрать ядерную функцию Гауссову или полиномиальную, после чего можно будет увидеть формулу по которой рассчитывается каждый элемент ядерной матрицы.

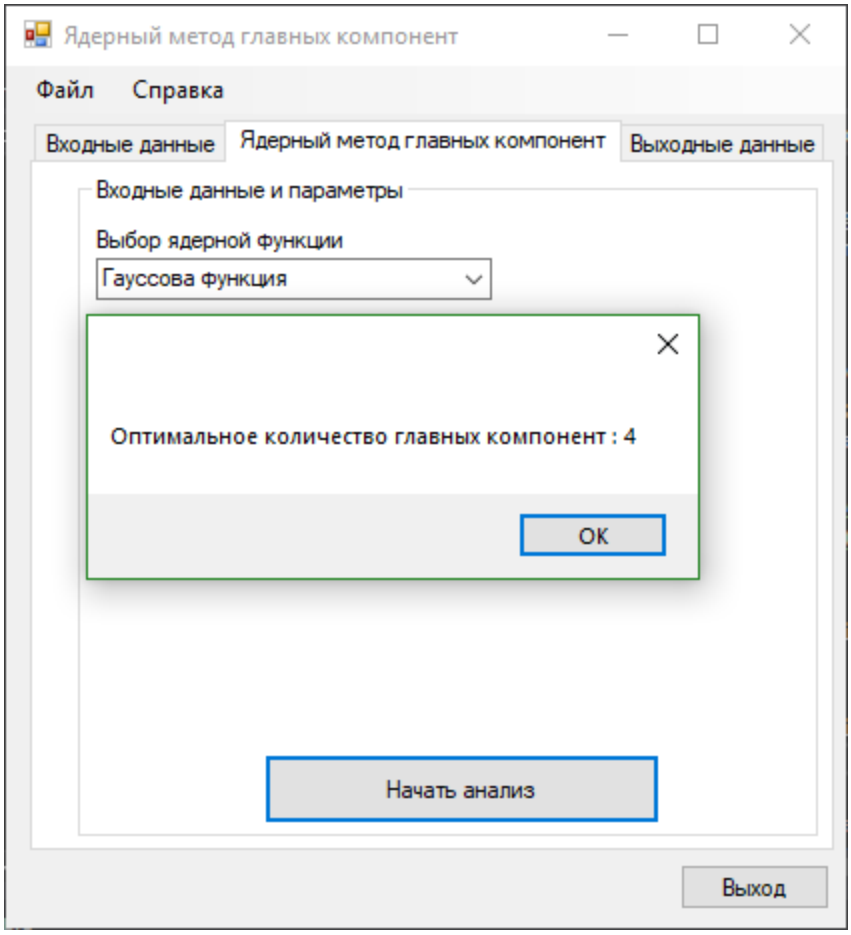

Рисунок 13 – Вкладка ядерный метод главных компонент основного окна

После нажатия на кнопку «Начать анализ» производятся все нужные вычисления для реализации ядерного метода главных компонент, выводит оптимальное количество главных компонент, алгоритм программы вычисляет ядерную матрицу на основе выбранной функции, центрирует её и диагонализирует, находит собственные значения и собственные вектора.

Далее пользователь может перейти на вкладку «Выходные данные» на рисунке 14, где можно увидеть выходные данные, полученные в результате работы алгоритма.

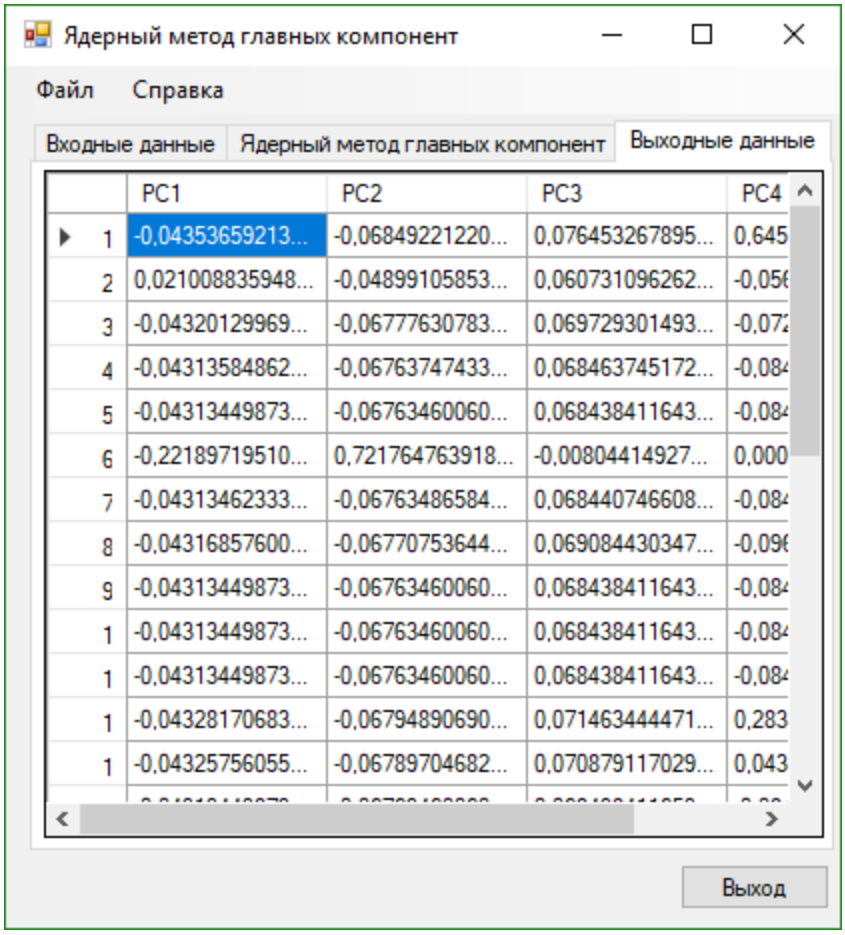

Рисунок 14 – Таблица выходных данных

Программа предоставляет возможность вывести входные и выходные данные и собственные значения на графики, окно для работы с графиками представлено на рисунке 15.

В окне «График» пользователь может выбрать какой график нужно построить, может выбрать какие главные компоненты на какую ось поместить, а также сохранить график в формате *.jpeg. Пример работы данного окна приведен на рисунке 15.

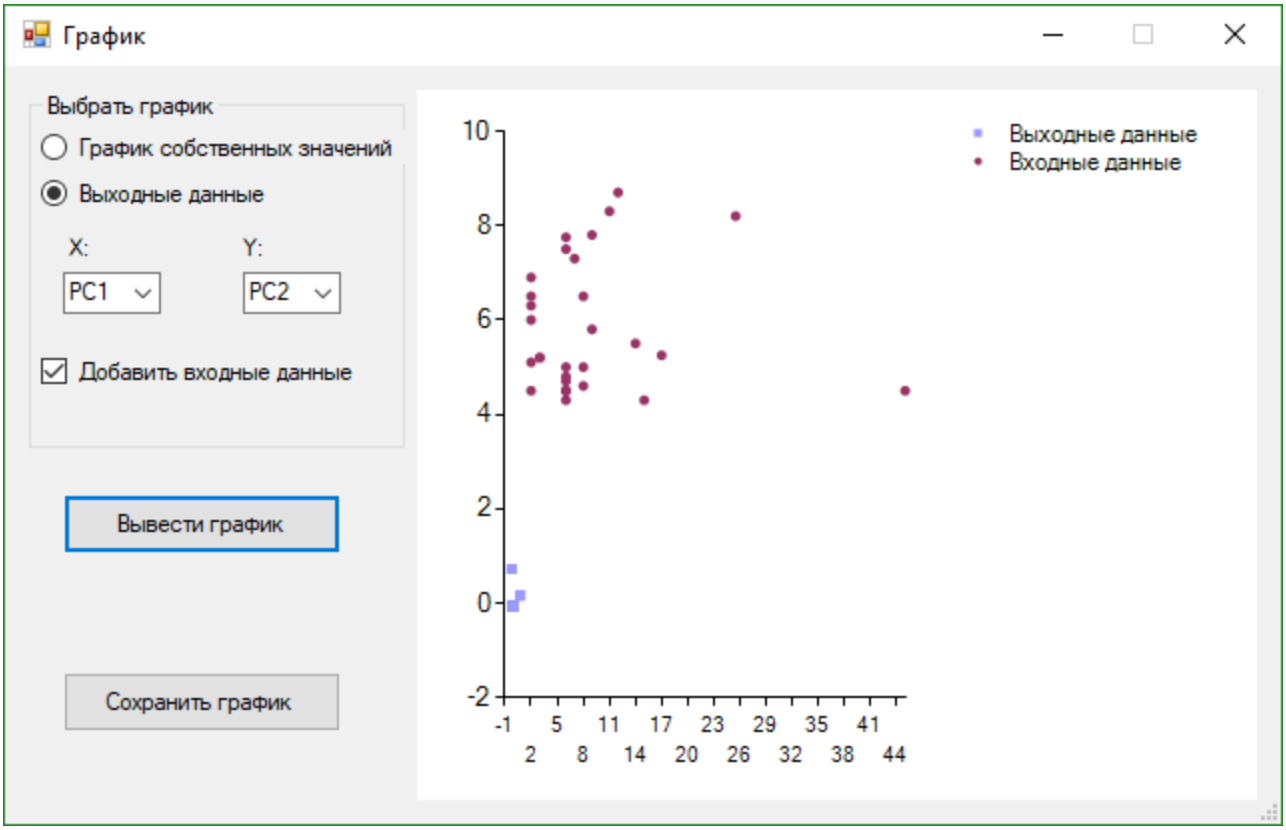

Рисунок 15 – Пример графика на основе входных и выходных данных

3 Пример использования программы

Разработанная программа может быть использована для анализа биомедицинских, экономических, информационных и других многомерных данных. Далее продемонстрируем результаты работы программы в виде файлов с данными и графиков. Имеется 2 набора данных определенного информационного эксперимента, уменьшим их размерность при помощи ядерного метода главных компонент.

Загрузим в программу файл «Данные1.csv», изображенный на рисунке 16 и применим ядерный метод главных компонент.

30	9							
6	4,8	1,2	3,41	0,54	11,8	2,43	22,38	0,03
2	6,9	3,45	1,19	2,29	13,1	2,55	10,7	0,07
2	6,3	1,89	2,18	3,02	10,1	1,7	19,1	0,08
12	8,7	4,7	3,1	1,15	7,05	1,35	7,8	0,14
45	4,5	1,76	2,73	0,7	5,7	1,63	18,8	0,04
6	4,7	2,12	2,46	1,46	5	3,16	18,8	0,09
9	7,8	3,82	2,3	1,89	17,3	3,37	14,6	0,06
6	7,75	2,95	0,7	0,54	7,2	1,8	7,85	0,11
25,5	8,2	2,46	1,19	1,62	13,1	5,68	43,1	0,1
8	4,6	2,99	2,05	1,2	8,95	8	75	0,1
11	8,3	3,74	1,67	1,35	8,6	9	49,3	0,11
8	6,5	2,86	1,29	0,78	10,95	3,6	24,3	0,13
2	4,5	2,25	2,62	1,35	13,1	2,11	21,4	0,61
6	5	2,1	2,32	1,59	11,1	2,53	7,21	9,21
3	5,2	2,24	2,62	0,94	12,4	1,82	8,56	0,08
2	6	2,04	2,99	1,62	13,1	1,36	9,04	0,13
15	4,3	1,72	1,16	1,75	5,7	2,99	18,5	0,1
9	5,8	3,07	2,03	3,97	6,3	2,08	6,57	0,16
6	7,5	2,25	1,19	1,46	6,26	2,42	23,89	0,12
2	5,1	2,04	1,92	0,98	5,35	2,53	12,6	0,08
8	5	1,85	1,55	1,21	18,2	2,21	9,87	0,15
2	6,5	1,95	3,89	2,29	13,1	1,28	7,7	32,36
3	5,2	2,65	0,8	2,02	8,5	1,64	7,01	1,94
6	4,5	1,26	2,89	1,91	8	4,99	28,7	0,11
6	4,5	1,4	3,89	1,48	6,8	3,1	9,7	0,06
6	4,3	1,68	1,89	1,11	6	2,48	17,5	0,12
17	5,25	1,73	2,05	1,06	3,89	2,3	18,1	0,07
6	4,5	2,57	4,32	0,92	12,2	3,18	26,1	0,19
14	5,5	1,93	1,04	1,02	6,78	2,6	21,3	0,15
7	7,3	2,7	1,04	1,21	5,2	2,9	13,4	0,19

Рисунок 16 – Входные данные файл «Данные1.csv»

PC1	PC2	PC3
-0,04354	-0,06849	0,076453
0,021009	-0,04899	0,060731
-0,0432	-0,06778	0,069729
-0,04314	-0,06764	0,068464
-0,04313	-0,06763	0,068438
-0,2219	0,721765	-0,00804
-0,04313	-0,06763	0,068441
-0,04317	-0,06771	0,069084
-0,04313	-0,06763	0,068438
-0,04313	-0,06763	0,068438
-0,04313	-0,06763	0,068438
-0,04328	-0,06795	0,071463
-0,04326	-0,0679	0,070879
-0,04313	-0,06763	0,068438
0,756761	0,163323	-0,01006
0,757321	0,161849	-0,00719
-0,04899	-0,0805	-0,68099
-0,04314	-0,06765	0,068547
-0,04314	-0,06764	0,068501
-0,04314	-0,06764	0,068488
-0,04313	-0,06763	0,068441
-0,04313	-0,06763	0,068438
-0,04294	-0,06763	0,068919
-0,04314	-0,06764	0,06846
-0,04314	-0,06765	0,068549
-0,2219	0,721762	-0,00804
-0,04735	-0,07703	-0,62241
-0,04343	-0,06826	0,074421
-0,04509	-0,07199	-0,25193
-0,04315	-0,06756	0,068466

Рисунок 17 – Выходные данные для «Данные1.csv»

Vector 1	Vector 2	Vector 3
-0,03368	-0,05634	0,074604
0,01625	-0,0403	0,059262
-0,03342	-0,05575	0,068042
-0,03337	-0,05564	0,066807
-0,03336	-0,05563	0,066783
-0,17164	0,593708	-0,00785
-0,03336	-0,05563	0,066785
-0,03339	-0,05569	0,067413
-0,03336	-0,05563	0,066783
-0,03336	-0,05563	0,066783
-0,03336	-0,05563	0,066783
-0,03348	-0,05589	0,069734
-0,03346	-0,05585	0,069164
-0,03336	-0,05563	0,066783
0,585353	0,134346	-0,00982
0,585787	0,133133	-0,00701
-0,03789	-0,06622	-0,66452
-0,03337	-0,05564	0,066889
-0,03337	-0,05564	0,066844
-0,03337	-0,05564	0,066831
-0,03336	-0,05563	0,066785
-0,03336	-0,05563	0,066783
-0,03322	-0,05563	0,067251
-0,03337	-0,05564	0,066804
-0,03337	-0,05564	0,06689
-0,17164	0,593705	-0,00784
-0,03662	-0,06336	-0,60735
-0,03359	-0,05615	0,072621
-0,03487	-0,05922	-0,24584
-0,03338	-0,05558	0,066809

Рисунок 18 – Собственные вектора.

На рисунке 17 изображены выходные данные, полученные через Гауссову функцию с тремя главными компонентами.

На рисунке 18 можно проанализировать собственные вектора полученные с помощью программы.

На рисунке 19 изображен график выходных и входных данных.

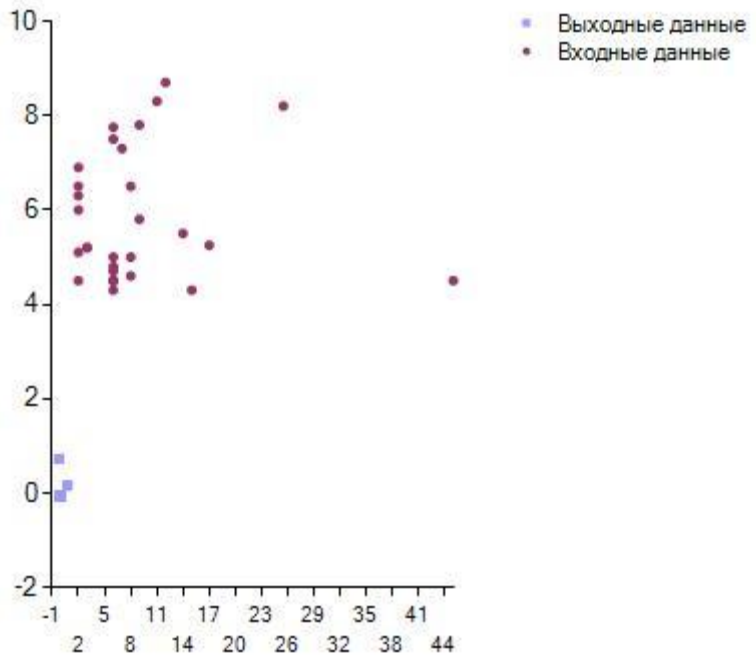

Рисунок 19 – График входных и выходных данных

На рисунке 20 изображен график собственных значений.

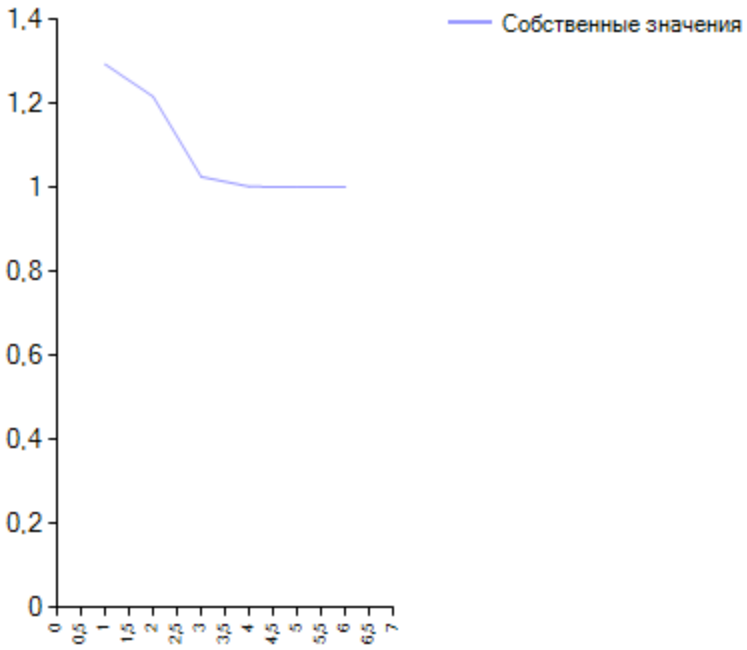

Рисунок 20 – График собственных значений

Теперь загрузим в программу файл «Данные2.csv», изображенные на рисунке 21, и применим ядерный метод главных компонент, проанализировав выходные данные на рисунке 22 и собственные вектора на рисунке 23.

30	9							
3	5	1,85	3,18	1,59	4,5	2,43	38,6	0,23
2	6	2,94	3,75	0,64	10,1	10,2	75	0,05
8	9,5	3,04	2,01	2,56	13,1	3,23	49,2	0,07
12	8,7	4,7	3,1	1,15	7,05	7,6	48	0,04
5	4,1	2,21	1,36	0,54	5,5	6,93	52,5	0,09
5	4,5	1,44	0,91	0,81	5,8	5,32	39,7	0,1
36	5,8	2,78	2,62	1,75	15,5	5,94	47,5	0,09
9	7,5	3,08	2,05	0,8	4,75	7,07	61,3	0,08
5	4,6	1,84	0,94	0,84	3,9	6,51	47,9	0,02
5	8,4	2,52	1,04	1,21	17,3	5,68	64,9	0,02
5	5,8	2,84	0,97	0,87	6,78	12	75	0,1
17	10	4,3	1,46	1,06	5,56	8,1	57,8	0,09
16,5	4,5	1,84	1,44	0,7	7,3	3,86	28,7	0,09
5	5	2,25	1,27	1,15	6,02	1,9	36,8	0,26
9	5,5	1,87	3,76	0,92	8	4,8	49,2	0,06
5	4	1,28	0,91	2,29	17,3	2,53	49,4	0,06
12	4,3	2,79	0,62	1,35	6,1	1,82	56	0,7
6	6,5	2,02	2,05	1,15	5,14	4,3	56,1	0,05
6	3,5	1,08	0,54	0,94	8	8,59	42,2	0,06
2	5,8	2,78	0,91	2,97	16,4	4,1	36	0,2
8	4,3	1,76	0,74	1,24	4,39	4	42,9	0,08
8	7,3	3,14	2,89	1,1	2,95	8,04	49,7	0,04
11	5	2,05	1,67	1,52	6,78	6,37	72,7	0,09
3	4,5	1,67	1,61	1,59	5,5	5,8	46,03	0,08
30	7	3,71	1,46	2,76	13,4	6,17	67,75	0,11
2	4,3	2,45	2,01	1,75	13,9	4,99	70,15	0,08
6	6,8	2,99	1,32	1,46	10,6	11,7	76	0,08
9	6,1	2,87	1,04	0,94	5,79	4,5	38,1	0,09
2	6,8	3,74	3,55	0,87	7,05	9,05	75	0,21
6	5,1	1,68	2,89	1,79	8,26	3,79	28,3	0,12

Рисунок 21 – Входные данные файл «Данные2.csv»

PC1	PC2	PC3
14,46127	-6,21665	1,866328
-22,8956	-6,21797	-1,05768
3,312988	0,370606	-5,28282
4,313024	3,313148	2,162525
0,166234	-4,15165	2,498812
13,03084	-4,1379	1,483978
5,048513	28,07616	-2,75646
-8,62112	-0,07403	4,147984
4,851981	-4,39278	3,780441
-12,73	-2,01569	-8,80614
-22,8627	-3,80873	2,828988
-5,32783	8,092049	4,692751
24,1592	7,416326	1,261653
16,2617	-4,00905	0,557378
3,551601	0,262443	0,243792
3,190741	-2,29272	-9,99813
-2,63468	2,964168	2,064075
-3,13479	-3,04208	2,576914
10,10964	-2,9068	0,157288
16,17804	-5,29292	-9,69664
10,13584	-1,3598	3,111608
2,827892	-1,39426	5,575893
-19,7835	2,107345	2,569705
6,683581	-6,11001	1,685484
-15,0551	22,0325	-1,13826
-17,5701	-5,674	-6,01103
-24,0424	-2,15456	-0,76661
14,67334	-0,05228	1,951351
-22,6452	-6,59765	1,667472
24,34658	-2,73323	-1,37065

Рисунок 22 – Выходные данные для «Данные2.csv»

Vector 1	Vector 2	Vector 3
0,002409	-0,00352	0,003764
-0,00381	-0,00352	-0,00213
0,000552	0,00021	-0,01065
0,000719	0,001878	0,004361
2,77E-05	-0,00235	0,005039
0,002171	-0,00235	0,002993
0,000841	0,015915	-0,00556
-0,00144	-4,20E-05	0,008365
0,000808	-0,00249	0,007624
-0,00212	-0,00114	-0,01776
-0,00381	-0,00216	0,005705
-0,00089	0,004587	0,009464
0,004025	0,004204	0,002544
0,002709	-0,00227	0,001124
0,000592	0,000149	0,000492
0,000532	-0,0013	-0,02016
-0,00044	0,00168	0,004162
-0,00052	-0,00172	0,005197
0,001684	-0,00165	0,000317
0,002695	-0,003	-0,01955
0,001689	-0,00077	0,006275
0,000471	-0,00079	0,011245
-0,0033	0,001195	0,005182
0,001114	-0,00346	0,003399
-0,00251	0,012489	-0,0023
-0,00293	-0,00322	-0,01212
-0,00401	-0,00122	-0,00155
0,002445	-2,96E-05	0,003935
-0,00377	-0,00374	0,003363
0,004056	-0,00155	-0,00276

Рисунок 23 – Собственные вектора для «Данные2.csv»

Далее проанализируем график выходных и входных данных на рисунке 24 и график собственных значений на рисунке 25.

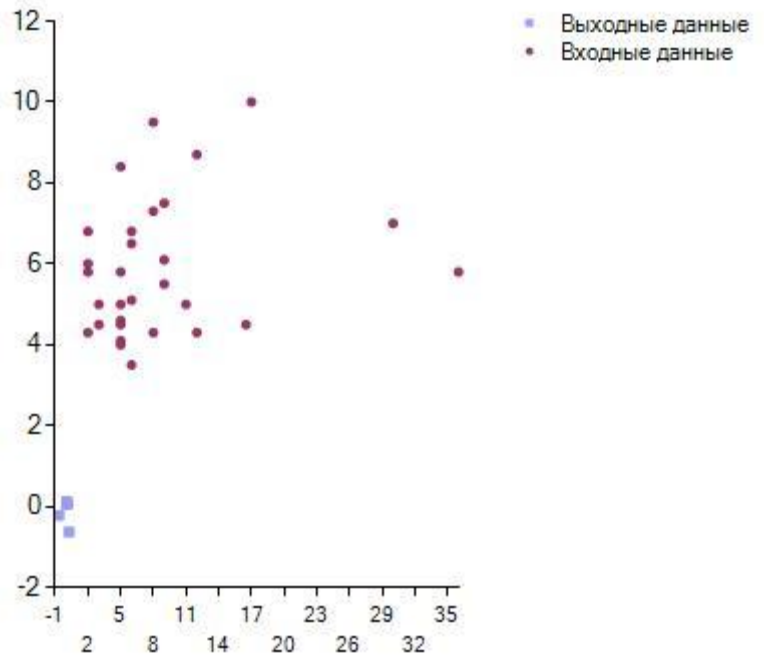

Рисунок 24 – График входных и выходных данных

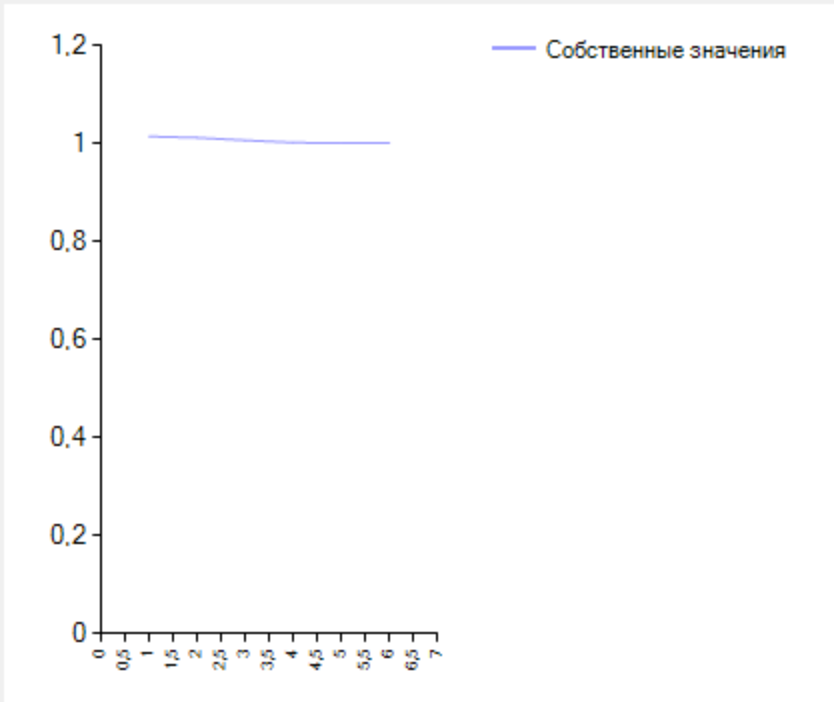

Рисунок 25 – График собственных значений

ЗАКЛЮЧЕНИЕ

В результате выполнения выпускной квалификационной работы было разработано программное обеспечение для анализа многомерных данных ядерным методом главных компонент. В данной пояснительной записке был описан процесс разработки программного обеспечения.

В процессе выполнения работы были выполнены следующие задачи:

– изучение классического и ядерного методов главных компонент;

– изучение существующих аналогичных программ обработки и анализа многомерных данных;

– выбор инструментария и способов решения;

– разработка программного обеспечения, реализующей ядерный метод главных компонент.

В результате работы была разработано программное обеспечение, позволяющее аналитику уменьшить размерность многомерных данных и визуализировать их при помощи ядерного метода главных компонент.

СПИСОК СОКРАЩЕНИЙ

МГК – метод главных компонент;

ЯМГК – ядерный метод главных компонент;

ВКР – выпускная квалификационная работа;

СФУ – Сибирский Федеральный Университет;

ПО – программное обеспечение.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Зиновьев, А. Ю. Визуализация многомерных данных / А. Ю. Зиновьев. – Красноярск: КГТУ, 2000. – 180 с.

2 MATLAB – высокоуровневый язык технических расчетов [Электронный ресурс]. – Режим доступа: https://matlab.ru/products/matlab/.

3 Smith, Lindsay I. A tutorial on Principal Components Analysis [Электронный ресурс]. – New York: Institute of Technology, 2002. – Режим доступа: https://www.ce.yildiz.edu.tr/personal/songul/file/1097/principal_components.pdf.

4 Van der Maaten, Laurens JP. Dimensionality reduction: A comparative review /Laurens JP Vander Maaten, Eric O Postma, H Jaap Van den Herik // Journal of Machine Learning Research. – 2009. – Т.10, – №1-41. – С.66-71.

5 Волкова, П. А. Статистическая обработка данных в учебно исследовательских работах / П. А. Волкова, А. Б. Шипунов. – Москва: ЭкоПресс-2000, 2008. – С.70-71.

6 Бых, А. И. Выбор метода восстановления пропущенных данных для оценки сердечно-сосудистой деятельности подростков / Бых А. И., Высоцкая Е.В.//Восточно-европейский журнал передовых технологий. – 2010. – Т.4, №45.

7 Shafranovich, Y. Common Format and MIME Type for Comma-Separated Values (CSV) Files [Электронный ресурс]. – IETF, 2005. – Режим доступа: http://www.ietf.org/rfc/rfc4180.txt.

8 Документация к пакету South [Электронный ресурс]. – Режим доступа: http://south.readthedocs.org/en/latest/.

9 Документация к пакету Psycopg 2 [Электронный ресурс]. – Режим доступа: http://initd.org/psycopg/.

10 Scikit-learn: Machine learning in Python / Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort [и др.] // The Journal of Machine Learning Research. – 2011. – Т.12. – С.2825-2830.